RESEARCH AND APPLICATION OF GIS AND DATA MINING TECHNOLOGY IN MONITORING AND ASSESSMENT OF NATURAL GEOGRAPHY ENVIRONMENT

X.L. ZHENG†

† Shanxi Coal Geological Geophysical Surveying Mapping Institute
TeresaAlicerCf@yahoo.com

***Abstract*— At present, with the acceleration of the economic development process, the maintenance of the ecological environment has received extensive attention. In order to simplify the workflow of natural geographical environment monitoring and evaluation, this paper combines GIS technology and data mining technology, and builds a decision tree model with monitoring and evaluation as the core. Dongting Lake is taken as the research object to verify the validity of the model. The research results show that the algorithm designed in this paper can classify the land types of natural geographical environment and improve the accuracy of environmental monitoring and evaluation.**

***Keywords* — GIS technology; data mining; natural geographical environment; monitoring and evaluation.**

I. INTRODUCTION

The condition of the natural geographical environment is related to the sustainable development of the economy and society and belongs to a complex ecosystem. Once a problem arises in the ecological environment, it will adversely affect the human's own survival and development. Therefore, in the process of using and transforming nature, it is necessary to conduct timely environmental monitoring and analysis, specify the specific contents of the ecological environment monitoring and evaluation index system and determine the appropriate monitoring and evaluation system based on the regional environment differences (Vale *et al.* 2016). Environmental monitoring is a process in which related departments conduct tests and monitoring on factors that affect environmental quality and development trends in accordance with relevant laws and regulations, thereby realizing the process of evaluating, supervising and controlling environmental quality (Fletcher-Lartey *et al., 2*016).

It includes the investigation of environmental background, research and development of monitoring programs, reasonable setting of monitoring sites, sample collection and inspection, testing of samples, collection of relevant data and comprehensive analysis (Liu *et al.* 2016). It is a scientific basis and an important guarantee for understanding the level of environmental quality, managing environmental pollution and strengthening environmental protection (Alanbari *et al.,* 2016). In short, environmental monitoring is an accurate, timely and comprehensive reflection of the current status and development trend of environmental quality, providing scientific basis for environmental management, pollution source control and environmental planning (Yamagata, 2017).

II.STATE OF THE ART

The current ubiquitous GIS system needs to complete the task of managing complex geographic data. At present, GIS technology is mainly focused on solving complex spatial data processing and display problems. The biggest difficulty encountered in its popularization and application is the lack of adequate thematic analysis models (Zheng, 2017). The data analysis ability of GIS is weaker, and the improvement of this ability fundamentally depends on the development and application of knowledge engineering, problem solving, planning, decision-making and automatic reasoning technology in artificial intelligence (Henriques *et al., 2*017). Faced with the ubiquitous situation of noise, persistence and non-linearity, which data mining method should be selected for the huge spatial data sets depends on the specific situation. It is often necessary to use data mining methods to solve the problems encountered in spatial data mining. In the monitoring of land, areas with an area of more than 100 km^2 can be easily measured with GIS satellites with an accuracy of up to 1%. Some scholars visually interpretate multi-phase data and dynamic monitoring of Harike land conditions in India based on Indian Remote Sensing Sate 1.1ite (IRS) images.

Scholars made comprehensive use of multi-temporal radars and NOAA and AVHRR remote sensing data from 1992 to 1998. Based on the establishment of band combinations and the effective extraction of water information, they analyzed and predicted the interannual and inter-seasonal water changes in Lake Chad, Africa. The law of flood variation in the river basin discussed the impact of the water movement in the Chad Basin on surrounding rivers, wetlands and swamps. They pointed out the intrinsic link between changes in the lake and hydrological conditions in the entire basin (Vajravelu *et al.*, 2018). The algorithm of data mining has the characteristics of accurate classification, which provides a very valuable technical support for the coming era of large data. At medium and low resolution, a large number of studies abroad have shown that the Landsat satellite data can effectively extract lake wetland contours, boundary shapes, the law of change and other important information.

III. METHODOLOGY

A. Texture Features of Gray Level Co-occurrence Matrix

In this paper, eight characteristics, including mean, variance, homogeneity, contrast, dissimilarity, information entropy, second-order matrix and correlation, were selected to calculate the texture features of remote sensing images. The calculation and description of each texture feature used are as follows: The mean value can be calculated as follows:

$$ME=\sum_{i=1}^{N}\sum_{j=1}^{N}(i-1)\times P(i,j) \quad (1)$$

N is the gray level of the remote sensing image. The mean feature can reflect the gray level uniformity in the texture window. The more regular the texture of the image, the larger the average value. The variance can be found as:

$$VA=\sum_{i=1}^{N}\sum_{j=1}^{N}(i-1)\times P(i,j)\times[(i-1)-ME]^{2} \quad (2)$$

The variance of texture features is a measure of the heterogeneity of the image, and the concept of variance in the statistical features of grayscale is similar. It also describes the degree of deviation from the overall mean of the sample. The smaller the image variance, the darker the variance feature image. The larger the variance, the brighter the variance feature image. The bright lines on this feature are mostly the edges of images in remote sensing images. Collaboration can be calculated as:

$$HO=\sum_{i=1}^{N}\sum_{j=1}^{N}P(i,j)/(1+(i-j)^{2}) \quad (3)$$

Collaboration can be used to reflect the local homogeneity of the image. The larger the image gray scale difference is, the smaller the cooperating value is, and it is expressed as a dark area on the feature image. The smaller the image difference is, that is, when the source image is a homogeneous area, the greater the cooperability value, the light area is represented as the bright area.

$$CON=\sum_{i=1}^{N}\sum_{j=1}^{N}P(i,j)\times(i-j)^{2} \quad (4)$$

Contrast can reflect the sharpness of the image and the texture of the grooves. It is the moment of inertia near the main diagonal of the GLCM. It can measure how the matrix values are distributed and the local changes in the image. The deeper the texture grooves are, the higher the contrast ratio is, and the clearer the effect is. The shallower the texture groove, the smaller the contrast value and the more blurred the effect.

$$DI=\sum_{i=1}^{N}\sum_{j=1}^{N}P(i,j)\times|(i-j)| \quad (5)$$

The dissimilarity reflects the difference in gray values between neighboring pixels. The smaller the difference in the gray value of the remote sensing image, the smaller the dissimilarity value. The greater the difference in the gray value, the greater the dissimilarity value.

$$ENT=\sum_{i=1}^{N}\sum_{j=1}^{N}-P(i,j)\times InP(i,j) \quad (6)$$

Information entropy can characterize the randomness of the texture of remote sensing images. It is mainly used to describe the disorder of remote sensing images in texture analysis. When the image contains many types of features or the image texture is messy, the gray distribution in the image is very random, and the value of the information entropy is larger.

$$SM=\sum_{i=1}^{N}\sum_{j=1}^{N}P(i,j)^{2} \quad (7)$$

The second-order moment is the sum of the squares of the elements of the gray level co-occurrence matrix and can be used to measure the change of the texture gray level of the remote sensing image. The second moment can reflect the uniformity of the image gray distribution and the degree of texture thickness. If the values in the texture window are consistent or have significant periodicity, the value is smaller. Conversely, if the distribution of values in the texture window is not uniform, the second moment is larger.

$$COR=\sum_{i=1}^{N}\sum_{j=1}^{N}P(i,j)\times(i-u_{x})/(\sigma_{x}\times\sigma_{y}) \quad (8)$$

Where：

$$\mu_{x}=\sum_{i=1}^{N}i\sum_{j=1}^{N}P(i,j)$$

$$\mu_{y}=\sum_{i=1}^{N}j\sum_{j=1}^{N}P(i,j)$$

$$\sigma_{x}=\sum_{i=1}^{N}(i-u_{x})^{2}\sum_{j=1}^{N}p(i,j)$$

$$\sigma_{y}=\sum_{j=1}^{N}(j-u_{y})^{2}\sum_{i=1}^{N}p(i,j)$$

The P(I, j) denotes the brightness value of the pixel in the *i*-th row and the *j*-th column position, and *N* denotes the size of the moving window when calculating the texture feature. The Eq. (8) can be used to describe the correlation between pixel gray values in remote sensing images. When there is a strong linear relationship between pixels, the texture correlation value is larger. When the correlation between pixels is small, the texture correlation value is also smaller. When similar texture areas in an image have a certain directionality, texture correlation values are also larger.

B. Selection of texture scales

The coarse texture is calculated with a large window, and the fine texture is calculated with a small window. Due to the higher resolution of SPOT-5 images, as the classification window increases, the classification time increases accordingly, which reduces the classification efficiency. Therefore, it is of great significance to improve the classification efficiency by selecting suitable texture features for different features. The step size of the texture feature texture of the SPOT-5 image panchromatic band was chosen to be 1.

The window size was 3×3, 5×5, 7×7 and 21×21. This paper uses the degree of separability of the fixed sample plots at different texture scales to determine the degree of separation of samples by each texture level. The separability test of the sample is usually used to determine whether the training area can be used as a training sample to further quantify the image. The detachability test of the training sample area mainly uses the Jeffries-Matusita and Transformed Divergence algorithm. This paper mainly uses the Jeffries-Matusita algorithm. The definition of the algorithm is described below:

$$JD(i,j) = 2*[1-\exp(-a(i,j))] \tag{9}$$

$$a(i,j) = 0.125 \times T[M(i)-M(j)] \times [A(i,j)] \times [M(i)-M(j)] + 0.5 \times \log[\frac{\det(A(i,j))}{\sqrt{\det(s(i) \times \det(s(i))}}] \tag{10}$$

$$A(i,j) = \frac{1}{2}[S(i)+S(j)] \tag{11}$$

Here $JD(i,j)$ represents the distance between categories i,j. $M(i),M(j)$ represents the mean vector of category i,j. The number of vector elements is the same as the number of image channels participating in the classification. $S(i),S(j)$ is the covariance matrix of the category, i,j, the number of rows and columns is the same as the number of images participating in the classification. The det() represents the determinant of the matrix value. From Eqn. (9) it can be seen that the result of *JD*(I, J) is between [0-2]. It represents the degree of separability between the selected regions of interest. Greater than 1.9 indicates that the selected scale has good separation for the sample, and if the value is low, the scale should be considered reasonable.

C. Decision Tree Model Construction

According to the statistics of different types of spectral features, through the trial and error of human-computer interaction, the best empirical value is selected as the threshold of each node in the decision tree using visual judgment methods. During the determination and adjustment of node thresholds, whether or not the selection of the threshold is reasonable is determined by comparing the degree of discrimination of the target ground class after the classification of different thresholds. For distinguishing between vegetation types and non-vegetation types, the normalized vegetation index (NDVI) of each pixel in the two scenic images of the study area in winter and summer is obtained by using band calculation.

Since Dongting Lake is characterized by rising water as a lake and falling into a river, the water level is low. After several comparisons of the winter image, the thresholds were set to 0. 22, that is, the pixels with winter image NDVI> 0. 22 were classified as vegetation. The woodland, TM3 (0.63 ~ 0.69 m) red band was the main absorption band of chlorophyll. For distinguishing plant types, coverage, and judging plant growth status. TM4 (0.76-0.90 m) is a near-infrared waveband. This waveband is located in a highly reflective area of a plant and reflects a large amount of plant information. It is used for plant identification and classification. The spectral characteristics of forest land (mainly poplar plantation) in the outer lake area are close to those of grass and beach in winter, and the spectrum of summer and reeds are close.

By contrast, the limited conditions can be set, that is, TM3<50, TM4<61 in winter images and TM3 <57, TM4 <66 in summer images. Due to limited space, related parameters such as mudflat, mossgrass, and woodland are not listed. Then we use the QUEST algorithm to train the training sample data (multispectral data and texture combination data of the best texture scale selected) to generate a decision tree.

IV. RESULTS AND DISCUSSION

A. Classification Results and Accuracy Verification

The above decision tree model was run to obtain the preliminary classification results of the SPOT-5 image. Then, the classification and post-processing were performed using recoding, clustering and removal analysis. The wetland classification map of the study area was obtained (Fig. 2).

This paper uses a random sampling method, reference resolution fusion image data combined with fusion image field survey. 30 sample points for each category, a total of 150 samples were selected.

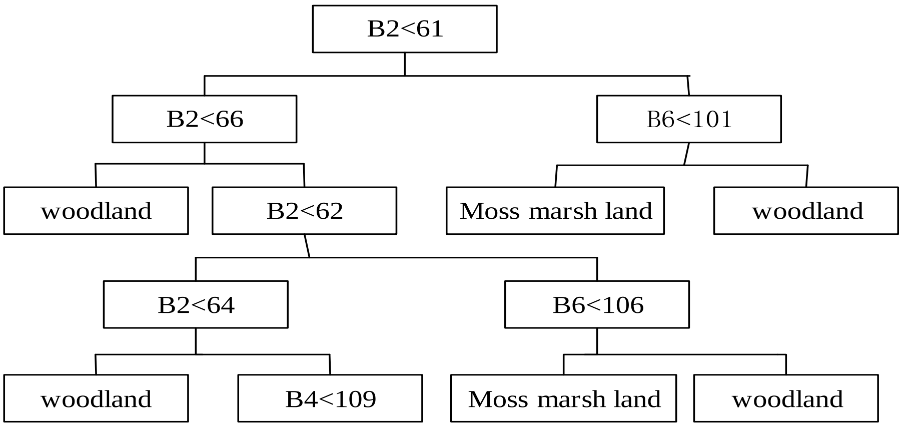

Figure1. Land Coverage Classification Based Texture by QUEST Decision Tree.

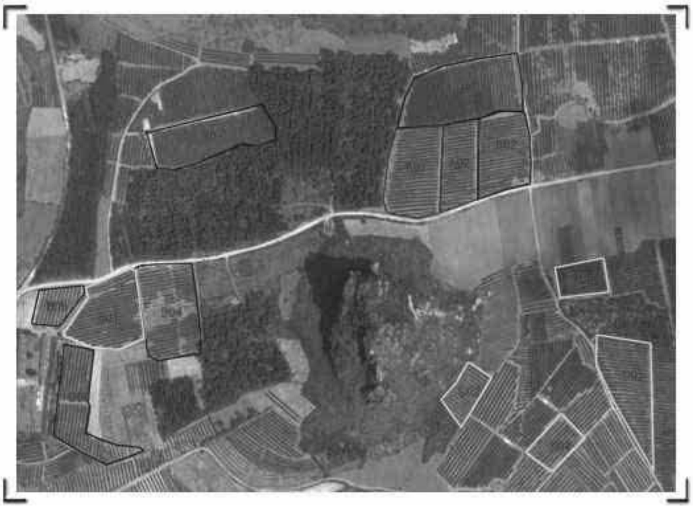

Figure 2. Classification using Multiscale Texture.

From Fig. 2, it can be seen that, all types of ground objects are basically patchy and less broken. Among them, the classification of water bodies, forest lands and reed beach lands is the best, and the accuracy of users reaches more than 90%. The producers of mudflats and moss-covered lands have the lowest precision, with the former being prone to confusion with water bodies and reeds, and the latter being prone to confusion with reed-flats. This is because mudflats are generally located in low-lying areas. Their soil moisture content is relatively high and they are easily confused with the spectrum of water bodies. For comparison, decision trees are classified separately for combining single-scale texture information and only multi-spectral data.

Finally, the classification results are obtained and the results are checked for accuracy. It is concluded that in general, various types of features are patchy and have a certain level of sense. Among them, water bodies and forest lands have the best classification effect, and user accuracy hits more than 90%. The accuracy of the producers of moss grass beach and reed beach is the lowest. Due to the lack of progress for moss grass beach and reed beach in the process of combining single-scale texture information, the textures of the two are relatively small, and the single-scale 17*17 window is used. Texture analysis generates a certain degree of confusion for small texture information, leading to a series of misclassifications.

Fig. 3 is the spectral information of the image. Fig. 4 is the classification situation. It can be seen from Fig. 4 that based on the multi-spectral classification information, its overall classification accuracy is much worse than the previous two because it ignores the texture information of the ground. The classification of forest lands is the best, and the user accuracy is only 89. 29% to 90%. The producer's accuracy in the reed land is only 72.97%.

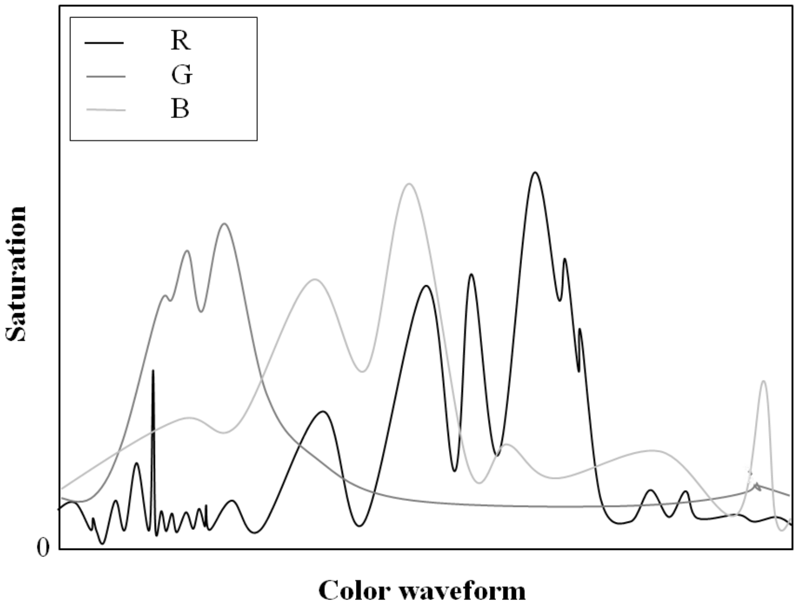

Figure 3. Image spectrum.

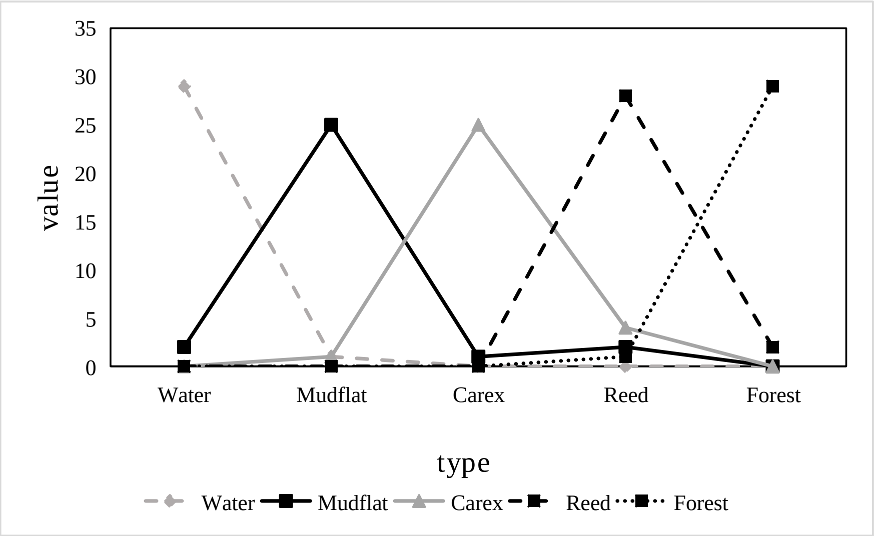

Figure4. Confusion Matrix of Classification Based on DTC Combined with Multispectral.

Through the aforementioned comparative analysis, we can see that the decision tree classification method based on comprehensive features such as spectrum and texture used in this paper has a greater improvement than the results obtained by simple spectral data in terms of single category accuracy, overall accuracy, or kappa coefficient. Higher accuracy can meet the requirements for accuracy of data acquisition in practical work. In the above section, by selecting the best combination of texture scales, high resolution image classification of spectral data and multi-scale texture data is performed using a decision tree.

The classification accuracy is 78.57%, while the classification accuracy of spectral data classification and single-scale texture data is 71.98 respectively. % and 76.76%. It can be seen that the multi-scale texture can better describe the texture features of the ground features and more effectively resolve the phenomenon of the same spectrum foreign objects in the classification results, which can help improve the high-resolution image classification accuracy.

V. CONCLUSIONS

In this paper, SPOT-5 high-resolution images were used to classify the land cover of Dongting Lake wetland, and the panchromatic band was selected as the data source for texture feature calculation. The JM distance of selected samples was used to determine the optimal texture scale corresponding to each wetland type. The decision tree algorithm performs data mining on the data set consisting of the spectrum and texture information of remote sensing images. A decision tree model was constructed and high-resolution images were classified. Results show that the classification accuracy is 78.57% with high-resolution image classification based on multi-scale optimal texture information, while the classification accuracy of single spectral data classification and single-scale texture data are 71.98% and 76.76% respectively. Therefore, the algorithm designed in this paper can contribute to the valuation of the geographical environment. It also influences positively the monitoring and evaluation of the natural environment.

REFERENCES

Alanbari, M.A., N. Alansari and H.K. Jasim, “GIS and Multicriteria Decision Analysis for Landfill Site Selection in Al-Hashimyah Qadaa”, *Natural Science*, **6(5)**, 282-304 (2016).

Fletcher-Lartey, S.M. and G. Caprarelli, “Application of GIS technology in public health, successes and challenges”, *Parasitology*, **143(4)**, 1-15 (2016).

Henriques, S, F. Guilhaumon and S. Villéger, “Biogeographical region and environmental conditions drive functional traits of estuarine fish assemblages worldwide”, *Fish & Fisheries*, **18(4)**, 752-771 (2017).

Liu, Y., X. Fang and C. Cheng, “Research and application of city ventilation assessments based on satellite data and GIS technology, a case study of the Yanqi Lake Eco-city in Huairou District, Beijing”, *Meteorological Applications*, **23(2)**, 320-327V (2016).

Vale, P.M, M.C.C. Stabile, “GIS without GPS, new opportunities in technology and survey research to link people and place”, *Population and Environment*, **37(4),** 391-410 (2016).

Yamagata, K., “Treatment of the natural environment in geography education and the role of Quaternary science”, *The Quaternary Research (Daiyonki-Kenkyu)*, **56(5),** 187-194 (2017).

Zheng, D. and D. Zhao, Characteristics of natural environment of the Tibetan Plateau”, *Science & Technology Review*, **35(6)**, 13-22 (2017).

Received: December 15th 2017
Accepted: June 30th 2018
Recommended by Guest Editor
Juan Luis García Guirao